Design Goal

The NCI Thesaurus is the reference vocabulary[1] for the National Cancer Institute.

Functionality

The NCI Thesaurus is intended to meet the internal needs of the Institute for comprehensive, controlled vocabulary. Because the activities of the Institute are diverse and extensive, any vocabulary adequate to meet the Institute's needs is likely to be of use to the broader cancer community. In that sense, the NCI Thesaurus may be adequate as a reference vocabulary for cancer research. Accordingly, the Thesaurus will be made freely available to the general public.

The principal intended use of the NCI Thesaurus within the Institute is to provide terminology for coding and retrieval functions. Coding and retrieval functions are commonly performed in connection with structured data (relational database table contents) and semi-structured data (markup tagged documents). The NCI Thesaurus is intended to facilitate operations with both types of data.

Coding and retrieval functions require that the NCI Thesaurus possess precise, logically correct and reliable hierarchical structure. Furthermore, the hierarchy of the Thesaurus must be navigable by users of our hierarchy browsing tools.

Design Principals

As a reference vocabulary, the NCI Thesaurus will contain, implicitly, all the concepts that NCI must be able to represent using its controlled vocabularies[2]. That is to say, the NCI Thesaurus will contain explicitly all the simple, or "atomic"[3], concepts needed to construct the complex terms and phrases that NCI requires.

The NCI Thesaurus is composed principally of atomic concepts. Concepts included are semantic entities required to:

1. represent the subject matter of cancer research,
2. qualify subject matter concepts, and
3. organize or aggregate subject matter concepts.

[1] "A reference terminology is "a set of concepts and relationships that provides a common reference point for comparison and aggregation of data about the entire health care process, recorded by multiple different individuals, systems, or institutions." 1 Excerpted from the SNOMED RT User Guide. Also referenced in Spackman KA, Campbell KE, Côté RA. SNOMED RT: a reference terminology for health care. Proceedings of the 1997 American Medical Informatics Association Fall Symposium. 1997.

[2] Controlled vocabulary is a set of terms or phrases authorized for use in the operations of an entity.

[3] An atomic concept is a semantic meaning that cannot be broken into constituent semantic units.

Inclusion or organization of concepts will be done as follows:

1. Complex concepts will be included in the NCI Thesaurus when they occur frequently in the NCI's work or when they have assumed *de facto* atomic status in common usage.
2. Concepts are organized into hierarchies using only is_a relationships. Right identity[4] is used to create part_of relationships.
3. Poly hierarchy[5] is permitted. The children of a concept must be valid at all positions within the hierarchy where the concept appears.
4. Concepts are placed into hierarchies algorithmically and manually. Manual placement is limited to placement of concepts that must be primitive (for example root concepts) or concepts in reference kinds[6].

Notes on Anticipated Use of NCI Thesaurus[7]

In applications that make use of the NCI Thesaurus will support two main user activities

1. Representation and recording of clinical, scientific and science management data, and
2. Retrieval, aggregation, and analysis of those data.

Reference properties are the properties of a reference terminology that support retrieval, aggregation and analysis. They provide a common reference point for the commonalities between different concepts. The primary reference property for all concepts in NCI Thesaurus is the "is_a" property which links concepts into a network (hierarchy) consisting of broad concepts at the very top of the hierarchy (e.g., Conceptual Entity, Process, Physical Entity, etc.) and increasingly more specific concepts (children) as one navigates down the hierarchy.

Many concepts in NCI Thesaurus are further defined by additional reference properties called Roles. Roles define semantic relationships between concepts. In NCI Thesaurusa for example, protein concepts are related to gene concepts by role "Encoded_By_Gene".

---

[4] Let A be a set on which there is a binary operation . An element e of this set is called a *right identity* if $a \bullet e = a$ for each a ε A

[5] In a poly hierarchy, concepts may occur in multiple trees (disjoint hierarchies).

[6] Reference kinds are hierarchies of concepts that are imported into a name space as a unit. The reference kind is always a primitive hierarchy, because it is not modeled in the name space, but is accepted from an outside authority as being valid. NCI Thesaurus employs reference kinds when well accepted vocabulary exists that is needed to model NCI concepts, and when the needed concepts are not central to the NCI's vocabulary. In NCI Thesaurus, for example, anatomy concepts are contained in a reference kind. The concept "hepatocellular carcinoma" could therefore refer to the anatomic concept "liver" through roles such as "occurs_in".

[7] This section adapted from the SNOMED RT User Guide and Spackman KA, Campbell KE, Côté RA. SNOMED RT: a reference terminology for health care. Proceedings of the 1997 American Medical Informatics Association Fall Symposium. 1997.

In contrast to reference properties, **interface properties** are the properties of a reference terminology that support representation and recording. They interact or interface with the user. These include the vocabulary and expressions presented to a user at the point of data entry. Interface properties may be required to create menu-based "pick lists," to provide acronyms, abbreviations, short phrases, and terms customized to a particular user or setting. Other characteristics that may be useful for data recording include optimizations for natural language processing, parts of speech, translations to and from foreign languages, and so forth.

Unlike SNOMED RT, NCI Thesaurus will focus on both the reference characteristics of terminology and on interface properties needed by NCI. Reference properties form the scaffolding or foundation on which the interface characteristics are built, and standardization of the reference characteristics can enable implementers and users to customize and optimize their own interfaces. However, because NCI Thesaurus is specialized to meet NCI's needs it is a priority to incorporate interface characteristics that make NCI Thesaurus easier for NCI application developers to use.

For example, NCI research clinicians may want to enter "AML - M2" instead of "Acute myelogenous leukemia, FAB M2." The former term is not in SNOMED, though the latter is. However both will be found in NCI Thesaurus.